Paper 72-28
Space odyssey: Concatenate zip files into one master file

Ying Long, Westat, Rockville, MD

ABSTRACT

There are many times when we receive national school data stored in multiple zip files. Separate school-level files must be concatenated into a single master file. The volume of national school data forces one to consider system resource constraints. This paper demonstrates how to minimize disk space requirements by using the SAS® System's X statement to execute host commands in order to get the names of zip files stored in a folder, unzip selected files, and then delete each file immediately after it has been processed. By taking advantage of the X command, one can avoid using WinZip, where you would have to unzip files individually and delete files manually. SAS is run in batch mode. The operating system we use is Windows 95. Base SAS & SAS macro facility are used in this paper. This paper is suited for all level of the SAS users.

INTRODUCTION

Although PCs are very powerful in terms of memory, speed, and disk space, there are still times when one has to deal with the disk space limitations on a project account. This is especially true when dealing with large national level survey data. The X command provides a way for a programmer to run DOS or Windows commands from within SAS. In some cases, this also provides opportunities for automation and system resource conservation.

TASK

Each zip file contains four files for a single school. We need to concatenate the file named S_<SCHID>FR2.SD7 to a master file. SCHID is a non-sequential school ID. There are approximately 250 zip files in the same directory.

STEPS

1. Create a text file that lists the name of each zip file in the directory.
2. Store the zip file name as a macro variable.
3. Unzip only the file we need to use.
4. Concatenate the file to the master file.
5. Delete the file to save space and speed up the process.

DISCUSSION

Let's take a closer look at the structure of the school level file using school 001 as an example.
S01_001F.ZIP contains 4 files as follows,

S_001FR2.SD7
S_001FRM.SD7
SMP_001.SD7
STAT_001.SD7

The file we need is S_001FR2.SD7.

Step 1

```
options noxwait xsync;
x "dir/o *.zip > outzip.txt";
```

The following shows some records from OUTZIP.TXT,

```
Volume in drive C is VOL1
 Directory of C:\My Documents\Paper

S01_001F ZIP    932,549  05-21-02  2:46p S01_001F.ZIP
S01_002F ZIP    743,867  05-21-02  2:47p S01_002F.ZIP
S01_008F ZIP    480,089  05-21-02  2:47p S01_008F.ZIP
S01_011F ZIP    610,865  05-21-02  2:47p S01_011F.ZIP
…
S01_116F ZIP    812,104  05-21-02  2:59p S01_116F.ZIP
S01_117F ZIP    669,067  05-21-02  2:59p S01_117F.ZIP
S01_122F ZIP    301,511  05-21-02  2:59p S01_122F.ZIP
…
S01_425F ZIP     30,130  05-21-02  3:12p S01_425F.ZIP
S01_426F ZIP    185,863  05-21-02  3:12p S01_426F.ZIP
S01_428F ZIP    541,059  05-21-02  3:12p S01_428F.ZIP
```

The SAS system option NOXWAIT specifies that user does not have to type EXIT to return control to SAS after the host command has executed. The SAS system option XSYNC specifies that control doesn't return to SAS until the host command has completed.

Step 2

```
data _null_;
length ii $3;
infile 'outzip.txt' missover;
input @1 dsn $8.;
if substr(dsn, 1, 1) ne 'S' then delete;
  i + 1;
```

```
    ii = left(put(i, 3.));
  call symput('n', ii);
  call symput('dsn'||ii, left(dsn));
  run;
```

The preceding data step assigns each file name and the number of files to a separate macro variable. Be careful that if your file name is longer than 8 characters, the TXT file output by the DIR command will truncate the name to 8 characters at the beginning of the line. Since DOS always gives out the full file name at the end of each line, you can easily get the file name by pointing your cursor to the right position.

Step 3

```
  data sas1.mast;
    if _n_ > 1;
  run;

  %macro unzip;

  %do j = 1 %to &n;
    %let sid = %substr(&&dsn&j, 5, 3);
    x "call pkunzip &&dsn&j..zip s_&sid.fr2.sd7";
```

First we create a blank master file, then we unzip the file we want from the zip file.

Step 4

```
  data sas1.mast;
    set sas1.mast sas1.s_&sid.fr2 (where=(_smpflag=.));
  run;
```

Then we select the records we want and concatenate them to the master file. Here SET is used because not all of the school-level files have the same variables, and we don't want to lose any information.

Step 5

```
  x "del s_&sid.fr2.sd7";
  %end;

  %mend unzip;
```

Finally, we delete the school-level file to save space. There is also a potential problem here when using the SET statement. The process will get slower and slower as the size of the master data set increases. To solve this problem, my solution is to concatenate a number of school-level files first, say 30 school-level files, then concatenate the intermediate files. Of course your computer will need sufficient memory and space to deal with this kind of problem.

An alternative way here is to use PROC APPEND instead of the SET statement. You'll have to create a master structure data file first. In order to do this, you need to get variable names of each of the files you are going to concatenate, get rid of the duplicated variables and set up a file with all the possible variables. PROC APPEND is supposed to be more efficient in concatenating files.

CONCLUSION

Working with large data sets, within the confines of limited disk space, may require a programmer to use non-SAS tools in addition to SAS to accomplish a task. The SAS System's X statement provides a convenient way to execute host commands to use the non-SAS tools.

REFERENCES

SAS Companion for the Microsoft Windows Environment. Version 6, 2nd Edition, SAS Institute Inc.

ACKNOWLEDGEMENTS

Many thanks to John Brown for helping me with the PKUNZIP part.
Many thanks to Randy Herbison for giving advice on this paper.

DISCLAIMER

The contents of this paper are the work of the author(s) and do not necessarily represent the opinions, recommendations, or practices of Westat.

CONTACT INFORMATION

Ying Long
Westat
1650 Research Boulevard
Rockville, MD 20850
Work Phone: (301)610-5584
Fax: (301)315-5934
Email: LongY1@Westat.com